

Package ‘STAARpipeline’

December 28, 2024

Type Package

Title STAARpipeline for Analyzing Whole-Genome/Whole-Exome Sequencing Data

Version 0.9.7.2

Date 2024-11-17

Author Xihao Li [aut, cre], Zilin Li [aut, cre], Sheila M. Gaynor [aut], Han Chen [aut]

Maintainer Xihao Li <xihao.li@unc.edu>, Zilin Li <lizl@nenu.edu.cn>

Description An R package for performing STAARpipeline in analyzing whole-genome/whole-exome sequencing data.

License GPL-3

Copyright See COPYRIGHTS for details.

Imports Rcpp, STAAR, MultiSTAAR, SCANG, dplyr, SeqArray, SeqVarTools, GenomicFeatures, TxDb.Hsapiens.UCSC.hg38.knownGene, GMMAT, GENESIS, Matrix, methods

Encoding UTF-8

LazyData true

Depends R (>= 3.2.0)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.2.3

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

Dynamic_Window_SCANG	2
fit_nullmodel	4
genesis2staar_nullmodel	6
Gene_Centric_Coding	7
Gene_Centric_Coding_cond	9
Gene_Centric_Coding_cond_spa	11
Gene_Centric_Noncoding	13
Gene_Centric_Noncoding_cond	15
Gene_Centric_Noncoding_cond_spa	17
Individual_Analysis	19
Individual_Analysis_cond	21
Individual_Analysis_cond_spa	22

LD_pruning	24
ncRNA	25
ncRNA_cond	27
ncRNA_cond_spa	29
Sliding_Window	31
Sliding_Window_cond	33
Sliding_Window_cond_spa	35
staar2scang_nullmodel	37

Index	39
--------------	-----------

Dynamic_Window_SCANG	<i>Genetic region analysis of dynamic windows using SCANG-STAAAR procedure</i>
----------------------	--

Description

The Dynamic_Window_SCANG function takes in chromosome, starting location, ending location, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between a quantitative/dichotomous phenotype and variants in a genetic region by using SCANG-STAAAR procedure. For each dynamic window, the scan statistic of SCANG-STAAAR-O is the set-based p-value of an omnibus test that aggregated p-values across different types of multiple annotation-weighted variant-set tests SKAT(1,1), SKAT(1,25), Burden(1,1) and Burden(1,25) using ACAT method; the scan statistic of SCANG-STAAAR-S is the set-based p-value of STAAAR-S, which is an omnibus test that aggregated p-values across multiple annotation-weighted variant-set tests SKAT(1,1) and SKAT(1,25) using ACAT method; the scan statistic of SCANG-STAAAR-B is the set-based p-value of STAAAR-B, which is an omnibus test that aggregated p-values across multiple annotation-weighted variant-set tests Burden(1,1) and Burden(1,25) using ACAT method.

Usage

```
Dynamic_Window_SCANG(
  chr,
  start_loc,
  end_loc,
  genofile,
  obj_nullmodel,
  Lmin = 40,
  Lmax = 300,
  steplength = 10,
  rare_maf_cutoff = 0.01,
  p_filter = 1e-08,
  f = 0,
  alpha = 0.1,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

chr	chromosome.
start_loc	starting location (position) of the genetic region to be analyzed using SCANG-STAAR procedure.
end_loc	ending location (position) of the genetic region to be analyzed using SCANG-STAAR procedure.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is the output from fit_nullmodel function and transformed using the staar2scang_nullmodel function.
Lmin	minimum number of variants in searching windows (default = 40).
Lmax	maximum number of variants in searching windows (default = 300).
steplength	difference of number of variants in searching windows, that is, the number of variants in searching windows are Lmin, Lmin+steplength, Lmin+steplength,..., Lmax (default = 10).
rare_maf_cutoff	a cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
p_filter	a filtering threshold of screening method for SKAT in SCANG-STAAR. SKAT p-values are calculated for regions whose p-value is possibly smaller than the filtering threshold (default = 1e-8).
f	an overlap fraction, which controls for the overlapping proportion of detected regions. For example, when f=0, the detected regions are non-overlapped with each other, and when f=1, we keep every susceptible region as detected regions (default = 0).
alpha	family-wise/genome-wide significance level (default = 0.1).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.
Use_annotation_weights	use annotations as weights or not (default = TRUE).
Annotation_name	a vector of annotation names used in SCANG-STAAR (default = NULL).
silent	logical: should the report of error messages be suppressed (default = FALSE).

Value

The function returns a list with the following members:

SCANG_O_res: A matrix that summarizes the significant region detected by SCANG-STAAR-O, including the negative log transformation of SCANG-STAAR-O p-value ("-logp"), chromosome

("chr"), start position ("start_pos"), end position ("end_pos"), family-wise/genome-wide error rate (GWER) and the number of variants ("SNV_num").

SCANG_O_top1: A vector of length 4 which summarizes the top 1 region detected by SCANG-STAAR-O. including the negative log transformation of SCANG-STAAR-O p-value ("-logp"), chromosome ("chr"), start position ("start_pos"), end position ("end_pos"), family-wise/genome-wide error rate (GWER) and the number of variants ("SNV_num").

SCANG_O_emthr: A vector of Monte Carlo simulation sample for generating the empirical threshold. The 1-alpha quantile of this vector is the empirical threshold.

SCANG_S_res, SCANG_S_top1, SCANG_S_emthr: Analysis results using SCANG-STAAR-S. Details see SCANG-STAAR-O.

SCANG_B_res, SCANG_B_top1, SCANG_B_emthr: Analysis results using SCANG-STAAR-B. Details see SCANG-STAAR-O.

References

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

Li, Z., Li, X., et al. (2019). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *The American Journal of Human Genetics*, 104(5), 802-814. ([pub](#))

Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

Liu, Y., et al. (2019). Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3), 410-421. ([pub](#))

fit_nullmodel

Fitting generalized linear mixed model with known relationship matrices under the null hypothesis.

Description

The `fit_nullmodel` function is a wrapper of the `glmmkin` function from the `GMMAT` package that fits a regression model under the null hypothesis, which provides the preliminary step for subsequent variant-set tests in whole-genome sequencing data analysis. See `glmmkin` for more details.

Usage

```
fit_nullmodel(
  fixed,
  data = parent.frame(),
  kins,
  use_sparse = NULL,
  use_SPA = FALSE,
  kins_cutoff = 0.022,
  id,
  random.slope = NULL,
  groups = NULL,
  family = binomial(link = "logit"),
```

```

method = "REML",
method.optim = "AI",
maxiter = 500,
tol = 1e-05,
taumin = 1e-05,
taumax = 1e+05,
tauregion = 10,
verbose = FALSE,
...
)

```

Arguments

fixed	an object of class formula (or one that can be coerced to that class): a symbolic description of the fixed effects model to be fitted. For multiple phenotype analysis, formula recognized by lm , such as <code>cbind(y1,y2,y3) ~ x1 + x2</code> , can be used in fixed as fixed effects.
data	a data frame or list (or object coercible by <code>as.data.frame</code> to a data frame) containing the variables in the model.
kins	a known positive semi-definite relationship matrix (e.g. kinship matrix in genetic association studies) or a list of known positive semi-definite relationship matrices. The rownames and colnames of these matrices must at least include all samples as specified in the <code>id</code> column of the data frame <code>data</code> . If <code>kins</code> is <code>NULL</code> , <code>fit_nullmodel</code> will switch to the generalized linear model with no random effects.
use_sparse	a logical switch of whether the provided dense <code>kins</code> matrix should be transformed to a sparse matrix (default = <code>NULL</code>).
use_SPA	a logical switch determines if the null model fitting occurs in an imbalanced case-control setting (default = <code>FALSE</code>).
kins_cutoff	the cutoff value for clustering samples to make the output matrix sparse block-diagonal (default = 0.022).
id	a column in the data frame <code>data</code> , indicating the id of samples. When there are duplicates in <code>id</code> , the data is assumed to be longitudinal with repeated measures.
random.slope	an optional column indicating the random slope for time effect used in a mixed effects model for longitudinal data. It must be included in the names of <code>data</code> . There must be duplicates in <code>id</code> and <code>method.optim</code> must be "AI" (default = <code>NULL</code>).
groups	an optional categorical variable indicating the groups used in a heteroscedastic linear mixed model (allowing residual variances in different groups to be different). This variable must be included in the names of <code>data</code> , and <code>family</code> must be "gaussian" and <code>method.optim</code> must be "AI" (default = <code>NULL</code>).
family	a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See family for details of family functions).
method	method of fitting the generalized linear mixed model. Either "REML" or "ML" (default = "REML").
method.optim	optimization method of fitting the generalized linear mixed model. Either "AI", "Brent" or "Nelder-Mead" (default = "AI").
maxiter	a positive integer specifying the maximum number of iterations when fitting the generalized linear mixed model (default = 500).

tol	a positive number specifying tolerance, the difference threshold for parameter estimates below which iterations should be stopped (default = 1e-5).
taumin	the lower bound of search space for the variance component parameter τ (default = 1e-5), used when method.optim = "Brent". See Details.
taumax	the upper bound of search space for the variance component parameter τ (default = 1e5), used when method.optim = "Brent". See Details.
tauregion	the number of search intervals for the REML or ML estimate of the variance component parameter τ (default = 10), used when method.optim = "Brent". See Details.
verbose	a logical switch for printing detailed information (parameter estimates in each iteration) for testing and debugging purpose (default = FALSE).
...	additional arguments that could be passed to glm .

Value

The function returns an object of the model fit from [glmkin](#) (obj_nullmodel), whether the samples are under imbalanced case-control design (obj_nullmodel\$use_SPA) and whether the kins matrix is sparse when fitting the null model. See [glmkin](#) for more details.

References

- Chen, H., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4), 653-666. ([pub](#))
- Chen, H., et al. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics*, 104(2), 260-274. ([pub](#))
- Chen, H. (2023). GMMAT: Generalized linear Mixed Model Association Tests Version 1.4.2. ([web](#))

genesis2staar_nullmodel

Transforming the null model object fitted using GENESIS to the null model object to be used for STAAR

Description

The genesis2staar_nullmodel function takes in the object from fitting the null model using the GENESIS package and transforms it to the object from fitting the null model to be used for STAAR procedure.

Usage

```
genesis2staar_nullmodel(obj_nullmodel_genesis, use_SPA = FALSE)
```

Arguments

- obj_nullmodel_genesis
an object from fitting the null model, which is the output from fitNullModel function in the GENESIS package.
- use_SPA
a logical switch determines if the null model fitting occurs in an imbalanced case-control setting (default = FALSE).

Value

An object from fitting the null model for related samples to be used for STAAR procedure, which is the output from `fit_nullmodel` function.

References

- Gogarten, S.M., Sofer, T., Chen, H., et al. (2019). Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, 35(24), 5346-5348. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))
- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

Gene_Centric_Coding	<i>Gene-centric analysis of coding functional categories using STAAR procedure</i>
---------------------	--

Description

The `Gene_Centric_Coding` function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between a quantitative/dichotomous phenotype (including imbalanced case-control design) and coding functional categories of a gene by using STAAR procedure. For each coding functional category, the STAAR-O p-value is a p-value from an omnibus test that aggregated SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Gene_Centric_Coding(
  chr,
  gene_name,
  category = c("all_categories", "plof", "plof_ds", "missense", "disruptive_missense",
    "synonymous", "ptv", "ptv_ds", "all_categories_incl_ptv"),
  genofile,
  obj_nullmodel,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
```

```

Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL,
SPA_p_filter = TRUE,
p_filter_cutoff = 0.05,
silent = FALSE
)

```

Arguments

<code>chr</code>	chromosome.
<code>gene_name</code>	name of the gene to be analyzed using STAAR procedure.
<code>category</code>	the coding functional category to be analyzed using STAAR procedure. Choices include <code>all_categories</code> , <code>plof</code> , <code>plof_ds</code> , <code>missense</code> , <code>disruptive_missense</code> , <code>synonymous</code> , <code>ptv</code> , <code>ptv_ds</code> , <code>all_categories_incl_ptv</code> (default = <code>all_categories</code>).
<code>genofile</code>	an object of opened annotated GDS (aGDS) file.
<code>obj_nullmodel</code>	an object from fitting the null model, which is either the output from <code>fit_nullmodel</code> function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the <code>genesis2staar_nullmodel</code> function.
<code>rare_maf_cutoff</code>	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
<code>rv_num_cutoff</code>	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
<code>rv_num_cutoff_max</code>	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
<code>rv_num_cutoff_max_prefilter</code>	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
<code>QC_label</code>	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
<code>variant_type</code>	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
<code>geno_missing_imputation</code>	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
<code>Annotation_dir</code>	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
<code>Annotation_name_catalog</code>	a data frame containing the name and the corresponding channel name in the aGDS file.
<code>Use_annotation_weights</code>	use annotations as weights or not (default = TRUE).
<code>Annotation_name</code>	a vector of annotation names used in STAAR (default = NULL).
<code>SPA_p_filter</code>	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).

p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).
silent	logical: should the report of error messages be suppressed (default = FALSE).

Value

A list of data frames containing the STAAR p-values (including STAAR-O or STAAR-B in imbalanced case-control setting) corresponding to each coding functional category of the given gene.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

Gene_Centric_Coding_cond

Gene-centric conditional analysis of coding functional categories using STAAR procedure

Description

The Gene_Centric_Coding_cond function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between a quantitative/dichotomous phenotype and coding functional categories of a gene by using STAAR procedure. For each coding functional category, the conditional STAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait conditional p-values (e.g. conditional MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Gene_Centric_Coding_cond(
  chr,
  gene_name,
  category = c("plof", "plof_ds", "missense", "disruptive_missense", "synonymous", "ptv",
    "ptv_ds"),
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  method_cond = c("optimal", "naive"),
```

```

QC_label = "annotation/filter",
variant_type = c("SNV", "Indel", "variant"),
geno_missing_imputation = c("mean", "minor"),
Annotation_dir = "annotation/info/FunctionalAnnotation",
Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL
)

```

Arguments

chr	chromosome.
gene_name	name of the gene to be analyzed using STAAR procedure.
category	the coding functional category to be analyzed using STAAR procedure. Choices include plof, plof_ds, missense, disruptive_missense, synonymous, ptv, ptv_ds (default = plof).
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from fitNullModel function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
known_loci	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
method_cond	a character value indicating the method for conditional analysis. optimal refers to regressing residuals from the null model on known_loci as well as all co-variables used in fitting the null model (fully adjusted) and taking the residuals; naive refers to regressing residuals from the null model on known_loci and taking the residuals (default = optimal).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.

Use_annotation_weights

use annotations as weights or not (default = TRUE).

Annotation_name

a vector of annotation names used in STAAR (default = NULL).

Value

A data frame containing the conditional STAAR p-values (including STAAR-O) corresponding to each coding functional category of the given gene.

References

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))

Gene_Centric_Coding_cond_spa

Gene-centric conditional analysis of coding functional categories using STAAR procedure for imbalance case-control setting

Description

The Gene_Centric_Coding_cond_spa function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between an imbalanced case-control phenotype and coding functional categories of a gene by using STAAR procedure. For each coding functional category, the conditional STAAR-B p-value is a p-value from an omnibus test that aggregated conditional Burden(1,25) and Burden(1,1), together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
Gene_Centric_Coding_cond_spa(
  chr,
  gene_name,
  category = c("plof", "plof_ds", "missense", "disruptive_missense", "synonymous", "ptv",
    "ptv_ds"),
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
```

```

geno_missing_imputation = c("mean", "minor"),
Annotation_dir = "annotation/info/FunctionalAnnotation",
Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL,
SPA_p_filter = FALSE,
p_filter_cutoff = 0.05
)

```

Arguments

chr	chromosome.
gene_name	name of the gene to be analyzed using STAAR procedure.
category	the coding functional category to be analyzed using STAAR procedure. Choices include plof, plof_ds, missense, disruptive_missense, synonymous, ptv, ptv_ds (default = plof).
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from fitNullModel function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
known_loci	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.
Use_annotation_weights	use annotations as weights or not (default = TRUE).
Annotation_name	a vector of annotation names used in STAAR (default = NULL).

`SPA_p_filter` logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = FALSE).

`p_filter_cutoff` threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).

Value

a data frame containing the conditional STAAR p-values (including STAAR-B) corresponding to each coding functional category of the given gene.

References

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))

Gene_Centric_Noncoding

Gene-centric analysis of noncoding functional categories using STAAR procedure

Description

The `Gene_Centric_Noncoding` function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between a quantitative/dichotomous phenotype (including imbalanced case-control design) and noncoding functional categories of a gene by using STAAR procedure. For each noncoding functional category, the STAAR-O p-value is a p-value from an omnibus test that aggregated SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Gene_Centric_Noncoding(
  chr,
  gene_name,
  category = c("all_categories", "downstream", "upstream", "UTR", "promoter_CAGE",
    "promoter_DHS", "enhancer_CAGE", "enhancer_DHS"),
  genofile,
```

```

obj_nullmodel,
rare_maf_cutoff = 0.01,
rv_num_cutoff = 2,
rv_num_cutoff_max = 1e+09,
rv_num_cutoff_max_prefilter = 1e+09,
QC_label = "annotation/filter",
variant_type = c("SNV", "Indel", "variant"),
geno_missing_imputation = c("mean", "minor"),
Annotation_dir = "annotation/info/FunctionalAnnotation",
Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL,
SPA_p_filter = TRUE,
p_filter_cutoff = 0.05,
silent = FALSE
)

```

Arguments

chr	chromosome.
gene_name	name of the gene to be analyzed using STAAR procedure.
category	the noncoding functional category to be analyzed using STAAR procedure. Choices include all_categories, downstream, upstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS (default = all_categories).
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from fitNullModel function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.

Use_annotation_weights	use annotations as weights or not (default = TRUE).
Annotation_name	a vector of annotation names used in STAAR (default = NULL).
SPA_p_filter	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).
p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).
silent	logical: should the report of error messages be suppressed (default = FALSE).

Value

A list of data frames containing the STAAR p-values (including STAAR-O or STAAR-B in imbalanced case-control setting) corresponding to each noncoding functional category of the given gene.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

Gene_Centric_Noncoding_cond

Gene-centric conditional analysis of noncoding functional categories using STAAR procedure

Description

The Gene_Centric_Noncoding_cond function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between a quantitative/dichotomous phenotype and noncoding functional categories of a gene by using STAAR procedure. For each noncoding functional category, the conditional STAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait conditional p-values (e.g. conditional MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Gene_Centric_Noncoding_cond(
  chr,
  gene_name,
  category = c("downstream", "upstream", "UTR", "promoter_CAGE", "promoter_DHS",
```

```

    "enhancer_CAGE", "enhancer_DHS"),
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  method_cond = c("optimal", "naive"),
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
  Annotation_name = NULL
)

```

Arguments

<code>chr</code>	chromosome.
<code>gene_name</code>	name of the gene to be analyzed using STAAR procedure.
<code>category</code>	the noncoding functional category to be analyzed using STAAR procedure. Choices include downstream, upstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS (default = downstream).
<code>genofile</code>	an object of opened annotated GDS (aGDS) file.
<code>obj_nullmodel</code>	an object from fitting the null model, which is either the output from <code>fit_nullmodel</code> function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the <code>genesis2staar_nullmodel</code> function.
<code>known_loci</code>	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
<code>rare_maf_cutoff</code>	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
<code>rv_num_cutoff</code>	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
<code>rv_num_cutoff_max</code>	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
<code>rv_num_cutoff_max_prefilter</code>	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
<code>method_cond</code>	a character value indicating the method for conditional analysis. <code>optimal</code> refers to regressing residuals from the null model on <code>known_loci</code> as well as all co-variables used in fitting the null model (fully adjusted) and taking the residuals; <code>naive</code> refers to regressing residuals from the null model on <code>known_loci</code> and taking the residuals (default = <code>optimal</code>).
<code>QC_label</code>	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").

variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.
Use_annotation_weights	use annotations as weights or not (default = TRUE).
Annotation_name	a vector of annotation names used in STAAR (default = NULL).

Value

A data frame containing the conditional STAAR p-values (including STAAR-O) corresponding to the noncoding functional category of the given gene.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))
- Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))

Gene_Centric_Noncoding_cond_spa

Gene-centric conditional analysis of noncoding functional categories using STAAR procedure for imbalance case-control setting

Description

The Gene_Centric_Noncoding_cond_spa function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between an imbalanced case-control phenotype and noncoding functional categories of a gene by using STAAR procedure. For each noncoding functional category, the conditional STAAR-B p-value is a p-value from an omnibus test that aggregated conditional Burden(1,25) and Burden(1,1), together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
Gene_Centric_Noncoding_cond_spa(
  chr,
  gene_name,
  category = c("downstream", "upstream", "UTR", "promoter_CAGE", "promoter_DHS",
    "enhancer_CAGE", "enhancer_DHS"),
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
  Annotation_name = NULL,
  SPA_p_filter = FALSE,
  p_filter_cutoff = 0.05
)
```

Arguments

chr	chromosome.
gene_name	name of the gene to be analyzed using STAAR procedure.
category	the noncoding functional category to be analyzed using STAAR procedure. Choices include downstream, upstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS (default = downstream).
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
known_loci	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").

variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.
Use_annotation_weights	use annotations as weights or not (default = TRUE).
Annotation_name	a vector of annotation names used in STAAR (default = NULL).
SPA_p_filter	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = FALSE).
p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).

Value

A data frame containing the conditional STAAR p-values (including STAAR-B) corresponding to the noncoding functional category of the given gene.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))
- Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))

Individual_Analysis	<i>Individual-variant analysis using score test</i>
---------------------	---

Description

The Individual_Analysis function takes in chromosome, starting location, ending location, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between a quantitative/dichotomous phenotype (including imbalanced case-control design) and each individual variant in a genetic region by using score test. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait score test p-values by leveraging the correlation structure between multiple phenotypes.

Usage

```
Individual_Analysis(
  chr,
  start_loc,
  end_loc,
  genofile,
  obj_nullmodel,
  mac_cutoff = 20,
  subset_variants_num = 5000,
  QC_label = "annotation/filter",
  variant_type = c("variant", "SNV", "Indel"),
  geno_missing_imputation = c("mean", "minor"),
  tol = .Machine$double.eps^0.25,
  max_iter = 1000,
  SPA_p_filter = TRUE,
  p_filter_cutoff = 0.05
)
```

Arguments

chr	chromosome.
start_loc	starting location (position) of the genetic region for each individual variant to be analyzed using score test.
end_loc	ending location (position) of the genetic region for each individual variant to be analyzed using score test.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
mac_cutoff	the cutoff of minimum minor allele count in defining individual variants (default = 20).
subset_variants_num	the number of variants to run per subset for each time (default = 5e3).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "variant", "SNV", or "Indel" (default = "variant").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
tol	a positive number specifying tolerance, the difference threshold for parameter estimates in saddlepoint approximation algorithm below which iterations should be stopped (default = ".Machine\$double.eps^0.25").
max_iter	a positive integer specifying the maximum number of iterations for applying the saddlepoint approximation algorithm (default = "1000").
SPA_p_filter	logical: are only the variants with a score-test-based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).
p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05)

Value

A data frame containing the score test p-value and the estimated effect size of the minor allele for each individual variant in the given genetic region. The first 4 columns correspond to chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT).

References

Chen, H., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4), 653-666. ([pub](#))

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

Individual_Analysis_cond

Individual-variant conditional analysis using score test

Description

The Individual_Analysis_cond function takes in the data frame of individual variants, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between a quantitative/dichotomous phenotype and each (significant) individual variant by using score test. For multiple phenotype analysis (obj_nullmodel\$n.pheno > 1), the results correspond to multi-trait conditional score test p-values by leveraging the correlation structure between multiple phenotypes.

Usage

```
Individual_Analysis_cond(
  chr,
  individual_results,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  method_cond = c("optimal", "naive"),
  QC_label = "annotation/filter",
  variant_type = c("variant", "SNV", "Indel"),
  geno_missing_imputation = c("mean", "minor"),
  geno_position_ascending = TRUE
)
```

Arguments

chr	chromosome.
individual_results	the data frame of (significant) individual variants for conditional analysis using score test. The first 4 columns should correspond to chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT).
genofile	an object of opened annotated GDS (aGDS) file.

obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
known_loci	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
method_cond	a character value indicating the method for conditional analysis. <code>optimal</code> refers to regressing residuals from the null model on <code>known_loci</code> as well as all co-variables used in fitting the null model (fully adjusted) and taking the residuals; <code>naive</code> refers to regressing residuals from the null model on <code>known_loci</code> and taking the residuals (default = <code>optimal</code>).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "variant", "SNV", or "Indel" (default = "variant").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
geno_position_ascending	logical: are the variant positions in ascending order in the GDS/aGDS file (default = TRUE).

Value

A data frame containing the conditional score test p-value and the estimated effect size of the minor allele for each (significant) individual variant in `individual_results`.

References

- Chen, H., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4), 653-666. ([pub](#))
- Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))
- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

Individual_Analysis_cond_spa

Individual-variant conditional analysis using score test for imbalance case-control setting

Description

The `Individual_Analysis_cond_spa` function takes in chromosome, starting location, ending location, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between an imbalanced case-control phenotype and each individual variant in a genetic region by using score test.

Usage

```
Individual_Analysis_cond_spa(
  chr,
  individual_results,
  genofile,
  obj_nullmodel,
  QC_label = "annotation/filter",
  variant_type = c("variant", "SNV", "Indel"),
  geno_missing_imputation = c("mean", "minor"),
  tol = .Machine$double.eps^0.25,
  max_iter = 1000,
  SPA_p_filter = FALSE,
  p_filter_cutoff = 0.05
)
```

Arguments

<code>chr</code>	chromosome.
<code>individual_results</code>	the data frame of (significant) individual variants for conditional analysis using score test. The first 4 columns should correspond to chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT).
<code>genofile</code>	an object of opened annotated GDS (aGDS) file.
<code>obj_nullmodel</code>	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
<code>QC_label</code>	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
<code>variant_type</code>	type of variant included in the analysis. Choices include "variant", "SNV", or "Indel" (default = "variant").
<code>geno_missing_imputation</code>	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
<code>tol</code>	a positive number specifying tolerance, the difference threshold for parameter estimates in saddlepoint approximation algorithm below which iterations should be stopped (default = ".Machine\$double.eps^0.25").
<code>max_iter</code>	a positive integer specifying the maximum number of iterations for applying the saddlepoint approximation algorithm (default = "1000").
<code>SPA_p_filter</code>	logical: are only the variants with a score-test-based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value (default = FALSE).
<code>p_filter_cutoff</code>	threshold for the p-value recalculation using the SPA method (default = 0.05)

Value

A data frame containing the score test p-value and the estimated effect size of the minor allele for each individual variant in the given genetic region. The first 4 columns correspond to chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT).

References

- Chen, H., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4), 653-666. ([pub](#))
- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

LD_pruning

Linkage disequilibrium (LD) pruning procedure

Description

The LD_pruning function takes in chromosome, the object of opened annotated GDS file, the object from fitting the null model, and a given list of variants to perform LD pruning among these variants in sequential conditional analysis by using score test. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait sequential conditional analysis by leveraging the correlation structure between multiple phenotypes.

Usage

```
LD_pruning(
  chr,
  genofile,
  obj_nullmodel,
  variants_list,
  maf_cutoff = 0.01,
  cond_p_thresh = 1e-04,
  method_cond = c("optimal", "naive"),
  QC_label = "annotation/filter",
  variant_type = c("variant", "SNV", "Indel"),
  geno_missing_imputation = c("mean", "minor"),
  geno_position_ascending = TRUE
)
```

Arguments

chr	chromosome.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from <code>fit_nullmodel</code> function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the <code>genesis2staar_nullmodel</code> function.
variants_list	the data frame of variants to be LD-pruned in sequential conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT).
maf_cutoff	the cutoff of minimum minor allele frequency in defining individual variants to be LD-pruned (default = 0.01).
cond_p_thresh	the cutoff of maximum conditional p-value allowed for variants to be kept in the LD-pruned list of variants (default = 1e-04).

method_cond	a character value indicating the method for conditional analysis. optimal refers to regressing residuals from the null model on known_loci as well as all co-variables used in fitting the null model (fully adjusted) and taking the residuals; naive refers to regressing residuals from the null model on known_loci and taking the residuals (default = optimal).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "variant", "SNV", or "Indel" (default = "variant").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
geno_position_ascending	logical: are the variant positions in ascending order in the GDS/aGDS file (default = TRUE).

Value

A data frame containing the list of LD-pruned variants in the given chromosome.

References

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

ncRNA	<i>Gene-centric analysis of long noncoding RNA (ncRNA) category using STAAR procedure</i>
-------	---

Description

The ncRNA function takes in chromosome, gene name, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between a quantitative/dichotomous phenotype (including imbalanced case-control design) and the exonic and splicing category of an ncRNA gene by using STAAR procedure. For each ncRNA category, the STAAR-O p-value is a p-value from an omnibus test that aggregated SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (obj_nullmodel\$n.pheno > 1), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
ncRNA(
  chr,
  gene_name,
  genofile,
  obj_nullmodel,
```

```

rare_maf_cutoff = 0.01,
rv_num_cutoff = 2,
rv_num_cutoff_max = 1e+09,
rv_num_cutoff_max_prefilter = 1e+09,
QC_label = "annotation/filter",
variant_type = c("SNV", "Indel", "variant"),
geno_missing_imputation = c("mean", "minor"),
Annotation_dir = "annotation/info/FunctionalAnnotation",
Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL,
SPA_p_filter = TRUE,
p_filter_cutoff = 0.05,
silent = FALSE
)

```

Arguments

chr	chromosome.
gene_name	name of the ncRNA gene to be analyzed using STAAR procedure.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.
Use_annotation_weights	use annotations as weights or not (default = TRUE).
Annotation_name	a vector of annotation names used in STAAR (default = NULL).

SPA_p_filter	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).
p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).
silent	logical: should the report of error messages be suppressed (default = FALSE).

Value

A data frame containing the STAAR p-values (including STAAR-O) corresponding to the exonic and splicing category of the given ncRNA gene.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

ncRNA_cond	<i>Gene-centric conditional analysis of long noncoding RNA (ncRNA) category using STAAR procedure</i>
------------	---

Description

The ncRNA_cond function takes in chromosome, gene name, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between a quantitative/dichotomous phenotype and the exonic and splicing category of an ncRNA gene by using STAAR procedure. For each ncRNA category, the conditional STAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait conditional p-values (e.g. conditional MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
ncRNA_cond(
  chr,
  gene_name,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
```

```

method_cond = c("optimal", "naive"),
QC_label = "annotation/filter",
variant_type = c("SNV", "Indel", "variant"),
geno_missing_imputation = c("mean", "minor"),
Annotation_dir = "annotation/info/FunctionalAnnotation",
Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL
)

```

Arguments

chr	chromosome.
gene_name	name of the ncRNA gene to be analyzed using STAAR procedure.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
known_loci	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (chr), position (pos), reference allele (ref), and alternative allele (alt) (default = NULL).
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
method_cond	a character value indicating the method for conditional analysis. <code>optimal</code> refers to regressing residuals from the null model on <code>known_loci</code> as well as all co-variables used in fitting the null model (fully adjusted) and taking the residuals; <code>naive</code> refers to regressing residuals from the null model on <code>known_loci</code> and taking the residuals (default = <code>optimal</code>).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.

Use_annotation_weights
use annotations as weights or not (default = TRUE).

Annotation_name
a vector of annotation names used in STAAR (default = NULL).

Value

A data frame containing the conditional STAAR p-values (including STAAR-O) corresponding to the exonic and splicing category of the given ncRNA gene.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))
- Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))

ncRNA_cond_spa	<i>Gene-centric conditional analysis of long noncoding RNA (ncRNA) category using STAAR procedure for imbalance case-control setting</i>
----------------	--

Description

The ncRNA_cond_spa function takes in chromosome, gene name, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between an imbalanced case-control phenotype and the exonic and splicing category of an ncRNA gene by using STAAR procedure. For each ncRNA category, the conditional STAAR-B p-value is a p-value from an omnibus test that aggregated conditional Burden(1,25) and Burden(1,1), together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
ncRNA_cond_spa(
  chr,
  gene_name,
  genofile,
  obj_nullmodel,
  known_loci,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
```

```

    Annotation_name = NULL,
    SPA_p_filter = FALSE,
    p_filter_cutoff = 0.05,
    silent = FALSE
)

```

Arguments

<code>chr</code>	chromosome.
<code>gene_name</code>	name of the ncRNA gene to be analyzed using STAAR procedure.
<code>genofile</code>	an object of opened annotated GDS (aGDS) file.
<code>obj_nullmodel</code>	an object from fitting the null model, which is either the output from <code>fit_nullmodel</code> function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the <code>genesis2staar_nullmodel</code> function.
<code>known_loci</code>	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
<code>rare_maf_cutoff</code>	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
<code>rv_num_cutoff</code>	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
<code>rv_num_cutoff_max</code>	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
<code>rv_num_cutoff_max_prefilter</code>	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
<code>QC_label</code>	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
<code>variant_type</code>	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
<code>geno_missing_imputation</code>	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
<code>Annotation_dir</code>	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
<code>Annotation_name_catalog</code>	a data frame containing the name and the corresponding channel name in the aGDS file.
<code>Use_annotation_weights</code>	use annotations as weights or not (default = TRUE).
<code>Annotation_name</code>	a vector of annotation names used in STAAR (default = NULL).
<code>SPA_p_filter</code>	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = FALSE).
<code>p_filter_cutoff</code>	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).
<code>silent</code>	logical: should the report of error messages be suppressed (default = FALSE).

Value

A data frame containing the STAAR p-values (including STAAR-O) corresponding to the exonic and splicing category of the given ncRNA gene.

References

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))

Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

Sliding_Window

Genetic region analysis of sliding windows using STAAR procedure

Description

The Sliding_Window function takes in chromosome, starting location, ending location, sliding window length, the object of opened annotated GDS file, and the object from fitting the null model to analyze the association between a quantitative/dichotomous phenotype (including imbalanced case-control design) and variants in a genetic region by using STAAR procedure. For each sliding window, the STAAR-O p-value is a p-value from an omnibus test that aggregated SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Sliding_Window(
  chr,
  start_loc,
  end_loc,
  sliding_window_length = 2000,
  type = c("single", "multiple"),
  genofile,
  obj_nullmodel,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
```

```

    Annotation_name = NULL,
    SPA_p_filter = TRUE,
    p_filter_cutoff = 0.05,
    silent = FALSE
)

```

Arguments

<code>chr</code>	chromosome.
<code>start_loc</code>	starting location (position) of the genetic region to be analyzed using STAAR procedure.
<code>end_loc</code>	ending location (position) of the genetic region to be analyzed using STAAR procedure.
<code>sliding_window_length</code>	the (fixed) length of the sliding window to be analyzed using STAAR procedure.
<code>type</code>	the type of sliding window to be analyzed using STAAR procedure. Choices include <code>single</code> , <code>multiple</code> (default = <code>single</code>).
<code>genofile</code>	an object of opened annotated GDS (aGDS) file.
<code>obj_nullmodel</code>	an object from fitting the null model, which is either the output from <code>fit_nullmodel</code> function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the <code>genesis2staar_nullmodel</code> function.
<code>rare_maf_cutoff</code>	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
<code>rv_num_cutoff</code>	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
<code>rv_num_cutoff_max</code>	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
<code>rv_num_cutoff_max_prefilter</code>	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
<code>QC_label</code>	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
<code>variant_type</code>	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
<code>geno_missing_imputation</code>	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
<code>Annotation_dir</code>	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
<code>Annotation_name_catalog</code>	a data frame containing the name and the corresponding channel name in the aGDS file.
<code>Use_annotation_weights</code>	use annotations as weights or not (default = TRUE).
<code>Annotation_name</code>	a vector of annotation names used in STAAR (default = NULL).

SPA_p_filter	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).
p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).
silent	logical: should the report of error messages be suppressed (default = FALSE).

Value

A data frame containing the STAAR p-values (including STAAR-O or STAAR-B in imbalanced case-control setting) corresponding to each sliding window in the given genetic region.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))

Sliding_Window_cond	<i>Genetic region conditional analysis of sliding windows using STAAR procedure</i>
---------------------	---

Description

The Sliding_Window_cond function takes in chromosome, starting location, ending location, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between a quantitative/dichotomous phenotype and variants in a genetic region by using STAAR procedure. For each sliding window, the conditional STAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait conditional p-values (e.g. conditional MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Sliding_Window_cond(
  chr,
  start_loc,
  end_loc,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
```

```

rv_num_cutoff_max_prefilter = 1e+09,
method_cond = c("optimal", "naive"),
QC_label = "annotation/filter",
variant_type = c("SNV", "Indel", "variant"),
geno_missing_imputation = c("mean", "minor"),
Annotation_dir = "annotation/info/FunctionalAnnotation",
Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL
)

```

Arguments

chr	chromosome.
start_loc	starting location (position) of the sliding window to be analyzed using STAAR procedure.
end_loc	ending location (position) of the sliding window to be analyzed using STAAR procedure.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
known_loci	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
method_cond	a character value indicating the method for conditional analysis. <code>optimal</code> refers to regressing residuals from the null model on <code>known_loci</code> as well as all co-variables used in fitting the null model (fully adjusted) and taking the residuals; <code>naive</code> refers to regressing residuals from the null model on <code>known_loci</code> and taking the residuals (default = <code>optimal</code>).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").

Annotation_name_catalog
a data frame containing the name and the corresponding channel name in the aGDS file.

Use_annotation_weights
use annotations as weights or not (default = TRUE).

Annotation_name
a vector of annotation names used in STAAR (default = NULL).

Value

A data frame containing the conditional STAAR p-values (including STAAR-O) corresponding to the sliding window in the given genetic region.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))
- Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))

Sliding_Window_cond_spa

Genetic region conditional analysis of sliding windows using STAAR procedure for imbalanced case-control setting

Description

The Sliding_Window_cond_spa function takes in chromosome, starting location, ending location, the object of opened annotated GDS file, the object from fitting the null model, and the set of known variants to be adjusted for in conditional analysis to analyze the conditional association between an imbalanced case-control phenotype and variants in a genetic region by using STAAR procedure. For each sliding window, the conditional STAAR-B p-value is a p-value from an omnibus test that aggregated conditional Burden(1,25) and Burden(1,1), together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
Sliding_Window_cond_spa(
  chr,
  start_loc,
  end_loc,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
```

```

QC_label = "annotation/filter",
variant_type = c("SNV", "Indel", "variant"),
geno_missing_imputation = c("mean", "minor"),
Annotation_dir = "annotation/info/FunctionalAnnotation",
Annotation_name_catalog,
Use_annotation_weights = c(TRUE, FALSE),
Annotation_name = NULL,
SPA_p_filter = FALSE,
p_filter_cutoff = 0.05
)

```

Arguments

<code>chr</code>	chromosome.
<code>start_loc</code>	starting location (position) of the sliding window to be analyzed using STAAR procedure.
<code>end_loc</code>	ending location (position) of the sliding window to be analyzed using STAAR procedure.
<code>genofile</code>	an object of opened annotated GDS (aGDS) file.
<code>obj_nullmodel</code>	an object from fitting the null model, which is either the output from <code>fit_nullmodel</code> function, or the output from <code>fitNullModel</code> function in the GENESIS package and transformed using the <code>genesis2staar_nullmodel</code> function.
<code>known_loci</code>	the data frame of variants to be adjusted for in conditional analysis and should contain 4 columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
<code>rare_maf_cutoff</code>	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
<code>rv_num_cutoff</code>	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
<code>rv_num_cutoff_max</code>	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
<code>rv_num_cutoff_max_prefilter</code>	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
<code>QC_label</code>	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
<code>variant_type</code>	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
<code>geno_missing_imputation</code>	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
<code>Annotation_dir</code>	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
<code>Annotation_name_catalog</code>	a data frame containing the name and the corresponding channel name in the aGDS file.
<code>Use_annotation_weights</code>	use annotations as weights or not (default = TRUE).

Annotation_name	a vector of annotation names used in STAAR (default = NULL).
SPA_p_filter	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = FALSE).
p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).

Value

A data frame containing the conditional STAAR p-values (including STAAR-B) corresponding to the sliding window in the given genetic region.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))
- Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3), 263-275. ([pub](#))

staar2scang_nullmodel	<i>Transforming the null model object fitted using STAAR to the null model object to be used for SCANG-STAAR</i>
-----------------------	--

Description

The staar2scang_nullmodel function takes in the object from fitting the null model and transforms it to the object from fitting the null model to be used for SCANG-STAAR procedure.

Usage

```
staar2scang_nullmodel(obj_nullmodel)
```

Arguments

obj_nullmodel	an object from fitting the null model, which is either the output from fit_nullmodel function, or the output from fitNullModel function in the GENESIS package and transformed using the genesis2staar_nullmodel function.
---------------	--

Value

An object from fitting the null model for related samples to be used for SCANG-STAAR procedure, which is the output from fit_null_glmkin_SCANG function for related samples in the SCANG package.

References

- Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12), 1599-1611. ([pub](#))
- Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969-983. ([pub](#))
- Li, Z., Li, X., et al. (2019). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *The American Journal of Human Genetics*, 104(5), 802-814. ([pub](#))

Index

Dynamic_Window_SCANG, [2](#)

family, [5](#)

fit_nullmodel, [3](#), [4](#), [7](#), [8](#), [10](#), [12](#), [14](#), [16](#), [18](#),
[20](#), [22–24](#), [26](#), [28](#), [30](#), [32](#), [34](#), [36](#), [37](#)

formula, [5](#)

Gene_Centric_Coding, [7](#)

Gene_Centric_Coding_cond, [9](#)

Gene_Centric_Coding_cond_spa, [11](#)

Gene_Centric_Noncoding, [13](#)

Gene_Centric_Noncoding_cond, [15](#)

Gene_Centric_Noncoding_cond_spa, [17](#)

genesis2staar_nullmodel, [6](#), [8](#), [10](#), [12](#), [14](#),
[16](#), [18](#), [20](#), [22–24](#), [26](#), [28](#), [30](#), [32](#), [34](#),
[36](#)

glm, [6](#)

glmmkin, [6](#)

Individual_Analysis, [19](#)

Individual_Analysis_cond, [21](#)

Individual_Analysis_cond_spa, [22](#)

LD_pruning, [24](#)

lm, [5](#)

ncRNA, [25](#)

ncRNA_cond, [27](#)

ncRNA_cond_spa, [29](#)

Sliding_Window, [31](#)

Sliding_Window_cond, [33](#)

Sliding_Window_cond_spa, [35](#)

staar2scang_nullmodel, [3](#), [37](#)